

Tilburg University

Are factor analytical techniques used appropriately in the validation of health status questionnaires?

de Vet, H.C.W.; Adèr, H.J.; Terwee, C.B.; Pouwer, F.

Published in:
Quality of Life Research

Publication date:
2005

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):

de Vet, H. C. W., Adèr, H. J., Terwee, C. B., & Pouwer, F. (2005). Are factor analytical techniques used appropriately in the validation of health status questionnaires? A systematic review on the quality of factor analysis of the SF-36. *Quality of Life Research*, 14(5), 1203-1218.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Are factor analytical techniques used appropriately in the validation of health status questionnaires? A systematic review on the quality of factor analysis of the SF-36

Henrica C.W. de Vet¹, Herman J. Adèr^{1,2}, Caroline B. Terwee¹ & François Pouwer^{1,3}

¹*Institute for Research in Extramural Medicine (E-mail: HCW.deVet@vumc.nl);* ²*Department of Clinical Epidemiology and Biostatistics;* ³*Department of Medical Psychology, VU University Medical Center, Amsterdam, The Netherlands*

Accepted in revised form 1 November 2004

Abstract

Factor analysis is widely used to evaluate whether questionnaire items can be grouped into clusters representing different dimensions of the construct under study. This review focuses on the appropriate use of factor analysis. The Medical Outcomes Study Short Form-36 (SF-36) is used as an example. Articles were systematically searched and assessed according to a number of criteria for appropriate use and reporting. Twenty-eight studies were identified: exploratory factor analysis was performed in 22 studies, confirmatory factor analysis was performed in five studies and in one study both were performed. Substantial shortcomings were found in the reporting and justification of the methods applied. In 15 of the 23 studies in which exploratory factor analysis was performed, confirmatory factor analysis would have been more appropriate. Cross-validation was rarely performed. Presentation of the results and conclusions was often incomplete. Some of our results are specific for the SF-36, but the finding that both the application and the reporting of factor analysis leaves much room for improvement probably applies to other health status questionnaires as well. Optimal reporting and justification of methods is crucial for correct interpretation of the results and verification of the conclusions. Our list of criteria may be useful for journal editors, reviewers and researchers who have to assess publications in which factor analysis is applied.

Key words: Confirmatory factor analysis, Exploratory factor analysis, Health status, Methodological quality, Principal component analysis, SF-36, Systematic review, Validation

Introduction

In psychosocial and medical sciences, constructs such as ‘anxiety’ or ‘quality of life’ are usually measured by means of multi-item health status questionnaires. It is important that the validity of these instruments is extensively tested before they are used. An important step in the validation of multi-item questionnaires is factor analysis. Factor analysis is a statistical technique which is designed to reveal whether or not the pattern of responses on a number of items can be explained by a smaller number of underlying factors [1–3]. The aim may be either pure data reduction,

assessment of the factor structure (dimensions) being measured by the questionnaire, or investigating whether the questionnaire shows the same dimensions across different groups (structural reliability).

When no clear-cut ideas about the factor structure (number of dimensions and their mutual associations) exist, the factor structure of an instrument can best be investigated by means of *exploratory* factor analysis. If prior hypotheses exist, based on theory or previous analyses, *confirmatory* factor analysis is more appropriate: it can be used to test whether the data fit a premeditated factor structure [1].

There are two forms of exploratory factor analysis: principal component analysis and common factor analysis [1, 2]. Principal component analysis aims to explain all variance in the data set; common factor analysis only explains the common variance of all items, and not the unique variance of the items. Whether this leads to different outcomes is still a subject of discussion [1, 2, 4]. According to Floyd and Widaman [1] principal component analysis is most appropriate for data reduction, making it possible to represent the data in a minimal number of dimensions (factors). This method is useful when the aim is to group a set of items in a few meaningful clusters for which summary indices can be calculated. Common factor analysis provides valuable insights into the multivariate structure of an instrument, by explaining the covariance and correlation structure among the measured items [1, 2]. It aims to identify a set of more general factors (latent variables) that explain the covariances among the measured variables. In common factor analysis, the factor loadings can be interpreted as the regression weights for predicting the items from the latent variables, also producing correlations among these [1]. In practice, both methods are used in similar situations.

Confirmatory factor analysis makes it possible to test whether a hypothesized factor structure of a questionnaire (based either on empirical data or on theory) is supported by actual data. Structural equation modeling techniques are used to test hypotheses about relationships between observed variables (items) and factors. This requires formal specification of a model to be estimated. The goodness-of-fit of the model to the data can be assessed by means of fit indices, while the estimated coefficients provide information about the associations between items and factors, and also between factors [1].

Factor analysis is a complex, but flexible analytical procedure. When reading studies on factor analysis, we encountered a wide variation in methods used and in the ways the results were presented and interpreted. Our own need for guidelines to judge and interpret results of factor analysis made us initiate the present study. The quality of results of factor analysis greatly depends on whether researchers comply with the assumptions underlying the methods and the principles that

have been developed for appropriate application. Explicitness in rationale and details of the methods used in every step of the procedure are necessary to help other researchers to interpret, verify and replicate the results. To the best of our knowledge, no critical assessment of the quality of factor analysis in the validation of health status questionnaires has yet been performed.

Since there is a wealth of multi-item questionnaires that can be used, we had to restrict our selection in some way. Options were restriction to a specific journal and evaluate all factor analyses that were done on multi-item health status questionnaires, or examining all factor analyses in relevant journals in a specific year (e.g. 2003), or choose one specific health status questionnaire. We choose for one, well known, health status questionnaire because of several advantages: only one instead of many different health status questionnaires and its background (available data about hypothesized structure) had to be described, reasonings and discussions by the authors can be judged better, and it facilitates comparability of the results [not subject of this manuscript]. We decided to restrict our analysis to the Medical Outcomes Study Short Form-36 [5]. The SF-36 is a generic, 36-item questionnaire that was designed in the 1980s to measure the concept of health status. It has been reported on in over 1000 publications [6]. Exploratory factor analyses have been used to determine the factor structure of the SF-36 [7–9]. The manual of this instrument [8] describes eight subscales (dimensions) that can be calculated from 35 of the 36 items (one item about self-reported health transition is not included in the scores): (1) physical functioning (10 items); (2) social functioning (2 items); (3) role limitations due to physical problems (4 items); (4) role limitations due to emotional problems (3 items); (5) mental health (5 items); (6) energy/vitality (4 items); (7) pain (2 items); and (8) general health perception (5 items). These 8 subscales were subjected to factor analysis which resulted in the development of a physical health subscale and a mental health subscale. This factor analysis based on the 8 subscales and not on the 35 individual items is called a second-order factor analysis. Ware and Gandek [10] present an excellent overview of the development and psychometric evaluation of the reliability and validity of the SF-36. In many studies the factor structure

of the SF-36 has been assessed when applying the questionnaire to different populations, with respect to language, culture or health status.

The objective of this systematic review is, primarily, to perform a critical assessment of the use of factor analytical techniques in the construct validity of health status questionnaires; the SF-36 is used as an example.

Methods

Searching for relevant studies

The electronic Medline (1966–August 2002) database was used to search for relevant studies. The following free-text words were used: ‘SF-36’ in combination with ‘factor analysis’, ‘factor analyses’, ‘factor analytical’, ‘principal component’ and ‘validity’, ‘exploratory’, ‘confirmatory’, ‘factor structure’ or ‘latent structure’. Additionally, the reference lists of all relevant studies were scanned.

Study selection

We included studies that evaluated the factor structure of the SF-36, using *exploratory* factor analysis (principal components analysis or common factor analysis) or *confirmatory* factor analysis, and were published in a peer-reviewed journal (English language). Studies that evaluated the first-order factor structure (of the 35 individual items) and studies that evaluated the second-order factor structure (of the 8 subscales) were selected. Studies that reported on the factor structure of shorter versions of the SF-36 (e.g. SF-12 [11]) were not included. Two researchers, FP and HdV, checked the titles and abstracts for eligibility. In case of doubt the full paper was examined.

Checklist to evaluate the methodological quality of factor analysis

We developed a checklist to score the methodological quality of the factor analysis in the identified studies. This checklist encompasses the choice and justification of the methods used, items on data quality, and detailed information on statistical entities. As it is difficult to apply the theoretical criteria to empirical studies, we scored and

discussed five studies in a pilot run, in order to compile detailed instructions on how to score the items. The list of methodological criteria is presented in Table 1.

A. Choice and justification of the method of factor analysis

1. Basic choice for exploratory and/or confirmatory factor analysis

As mentioned in the introduction, it is important to make a distinction between exploratory factor analysis and confirmatory factor analysis, which may provide answers to different research questions, and pose different requirements on the data collected. The two techniques have different assumptions about the data and how they should be handled. For an elaborate account of exploratory factor analysis, we refer to Zwick and Velicer [12]. In short, principal component analysis or common factor analysis is applied to a set of items, and this results in an (unrotated) factor structure. To determine how many factors are meaningful, a number of methods are available [13]. The factors are rotated to facilitate interpretation. Orthogonal (e.g. Varimax) rotation, which is often used, assumes the factors to be uncorrelated. Oblique rotation allows for correlations between the factors [1]. The requirements for confirmatory factor analysis have been extensively described by Bollen [14]. The information to be considered includes the estimates of the parameters and their standard errors, and the appropriate goodness-of-fit measures.

The choice of the method of factor analysis (exploratory or confirmatory) is crucial. We assessed whether the choice of the specific method was appropriate for the research question that was addressed. We considered exploratory factor analyses to be appropriate if the aim of the study (according to the authors) was to examine the factor structure of the SF-36 in a patient population or language in which the SF-36 had not yet been used, without a prior hypothesis. If the aim of the study (according to the authors) was to confirm the existing first-order 8-factor structure or the second order 2-factor structure, we considered confirmatory factor analysis to be more appropriate. If exploratory factor analysis was used, it should be stated whether principal component analysis or common factor analysis was performed.

Table 1. Checklist and results

Item	Description	Scoring				
		+	−	?	0	N.A.
<i>A. Choice and justification of methods</i>						
1. Exploratory vs. confirmatory factor analysis						
	1.1 Is the type of factor analysis appropriate in view of the research question?	13	15	0	0	0
	1.2 When both types of factor analysis were used, has this analysis strategy been convincingly justified?	1	0	0	0	27
2. Exploratory factor analysis						
	2.1 Has the number of factors to be rotated been justified?	20	3	0	0	5
	2.2 Has the choice of the rotation method been justified?	2	1	0	0	25
	2.3 Is the interpretation of the final factor solution properly justified?	12	8	0	0	8
	2.4 In the case of a non-orthogonal factor structure, has the association between factors been discussed?	2	1	0	0	25
3. Confirmatory factor analysis						
	3.1 Has the model to be confirmed been well described?	6	0	0	0	22
	3.2 Has the strategy to arrive at the 'best' model been well described?	6	0	0	0	22
	3.3 Were the analysis results properly interpreted?	5	1	0	0	22
	3.4 Has the association between factors been discussed?	3	3	0	0	22
4. Cross-validation						
	4.1 Has cross-validation been applied in case this was possible?	4	21			3
	4.2 Has cross-validation been performed with different randomly drawn samples?		4			24
	4.3 If applied, did the number of observations justify this procedure?	4				24
	4.4 If applied, was the interpretation of the results convincing?	3	1			24
<i>B. Sample size and data quality</i>						
1. Sample size						
	1.1 Has the number of observations been sufficient to justify the use of factor analysis? ¹	25	3	0	0	0
	1.2 Has the number of observations been sufficient to perform cross-validation? ²	25	3	0	0	0
2. Data quality: missing data procedures						
	2.1 Does the study report on the percentage of missings?	15	13	0	0	0
	2.2 If this percentage is alarming (> 25%), is there information about whether the missing were considered random?	0	1	0	0	27
	2.3 If missing data have been imputed, was the imputation method appropriate?	8	0	0	0	20
3. Data quality: distributional properties						
	3.1 Have the distributional properties (at least standard deviations in exploratory factor analysis and kurtosis in confirmatory factor analysis) of the variables been reported?	21	7	0	0	0
	3.2 In case of undesirable distributional properties (lack of variance in exploratory factor analysis and excessive kurtosis in confirmatory factor analysis), have they been properly handled?	1	3	0	0	24
<i>C. Full report of statistical entities</i>						
1. Exploratory factor analysis (N = 23)						
	1.1 Principal component analyses or common factor analyses	Yes	No			NA
	1.2 Criteria for retaining factors	23	0			5
	1.3 Eigenvalues, percentage of variance accounted for by the (un)rotated factors	16	7			5
	1.4 Rotation method	14	9			5
	1.5 Rationale for rotation in case of oblique solutions	21	2			5
	1.6 All rotated factor loadings	2	1			25
	1.7 Factor inter-correlation in oblique solutions	20	3			5
		2	1			25
2. Confirmatory factor analysis (n = 6)						
	2.1 Number of factors	6	0			22
	2.2 Composition of factors	6	0			22
	2.3 Orthogonal vs. correlated factors	4	2			22
	2.4 Other model constraints (fixed and free parameters)	3	3			22
	2.5 Methods of estimation	5	1			22
	2.6 Overall fit	6	0			22

Table 1. *Continued*

Item	Description	Scoring				
		+	–	?	0	N.A.
	2.7 Relative fit	4		2		22
	2.8 Parsimony	2				26
	2.9 Any model modification to improve model fit to data	5		0		23
	2.10 Factor loadings	4		2		22
	2.11 Communalities (or squared correlations of observed variables with the factors)	1		5		22
	2.12 Factor correlations	3		3		22

* – Number of scores, for papers for which the question was relevant.

+ – The description is both informative and methodologically sound.

– – The description is informative, but methodologically doubtful.

? – The description is too unclear or too incomplete to answer the question.

0 – No relevant information about this item is given in the paper.

¹Smallest subsample was scored.

²Largest subsample was scored.

2. Performance of exploratory factor analysis

Justification of the choice of methods within exploratory analysis is required. This includes (1) a justification of the number of factors to be rotated: authors should have used and described at least one of the criteria for retaining factors before rotation; (2) a justification of the choice of the rotation method (if orthogonal (e.g. Varimax) rotation was used, no justification was required); (3) an elucidation of the interpretation of the factor solution and a discussion of the consequences: these should include a final choice for a specific factor structure and a justification of this choice; and (4) if factor analysis with non-orthogonal rotation was used, the associations between factors should be discussed.

3. Performance of confirmatory factor analysis

This is based on a hypothesized factor structure (model). The following quality criteria are important: (1) the factor structure to be confirmed should be clearly described; (2) the strategy that was applied to arrive at the best model should be clearly described; (3) the results of the analyses should be correctly interpreted; and (4) the associations between factors should be discussed.

4. Performance of cross-validation

Cross-validation means that the results of factor analysis of one part of the data set are tested on another part of the same data set. To ascertain that these two parts are as comparable as possible,

it is essential that the two subsamples are selected at random. Cross-validation gives an indication of the stability of the factor structure, and it is a requirement for both exploratory and confirmatory factor analysis. Obviously, splitting the data randomly into two subsamples requires twice the usual sample size. A potentially useful practice for cross-validation is to perform exploratory factor analysis on one half of the data set and then perform confirmatory factor analysis on the other half to confirm the factor structure [15].

B. Sample size and data Quality

1. Sample size

The number of subjects included in the analysis is another point that must be considered. Both exploratory factor analysis and confirmatory factor analysis require a reasonable amount of data to produce reliable results. The debate on the minimal requirement is still ongoing [1]. Rules-of-thumb vary from 4 to 10 subjects per variable, with a minimum number of 100 subjects to ensure stability of the variance-covariance matrix [16]. The required number of subjects per variable depends, among other things, on the factor loadings, on the number of variables per factor [17], and on the total sample size, with a smaller sample size requiring more subjects per variable [3]. We decided to set the requirement to 7 subjects per item. For the SF-36 this means that in a first-order factor analysis (focusing on the 35 individual items), at least 245

subjects should be available. For a second-order analysis on the 8 subscale scores, 56 subjects would be sufficient, but in that case we applied the rule-of-thumb of a minimum of 100 subjects.

2. Missing data

It is important to provide information about the number of missing values on each item, and how the missing values were dealt with. For, an accumulation of missing values due to paired missings may seriously curtail the number of subjects on which the variance–covariance matrix is based. Missing values of more than 25% in any item was considered to be the maximum, in which case, information should be provided as to whether missing values were considered to be random. If the percentage of missing values was acceptable (<25%) or if these were considered to be missing at random, imputation of mean values was scored as an acceptable method, but complete case analysis was considered to be inadequate.

3. Distributional concerns

Exploratory factor analysis requires a normal distribution of the data. Confirmatory factor analysis requires sufficient variation in the data. However, only substantial violations of normality or lack of variance, respectively, affect the results. Nevertheless, we assessed whether the authors described the distribution (means and standard deviations) of the data, and how they dealt with any violations of normality or lack of variance.

C. Full report of statistical entities

Floyd and Widaman [1] have developed a checklist of statistical entities that should be reported in order to enable a correct interpretation of the results. The Floyd and Widaman checklist focuses mainly on the technical aspects of the procedures. They explained the importance and interpretation of all items in detail [1].

Methodological assessment of the papers

Two members of our review group (HA and CT) independently evaluated the quality of the studies according to the afore-mentioned checklist. Disagreements were resolved by consensus, with or without a third reviewer (HdV).

Results

The PubMed search yielded 83 publications. Based on the abstracts, 34 studies were likely to meet the inclusion criteria. Full reports were then obtained from the library. Of the 34 studies 8 did not meet our eligibility criteria for the following reasons: six studies performed no factor analysis [10, 18–22], and in two studies the factor analysis was not performed on the SF-36 [23, 24]. A new search by an independent reviewer yielded one additional study [25] and another study [26] was identified by correspondence with authors (both studies were not available in PubMed). So, two eligible studies were included during the reviewing process [25, 26].

Table 2 presents an overview of the characteristics of the 28 included studies. Only six studies performed a first-order factor analysis on the 35 items of the SF-36 [27–32]. Twenty-five studies performed a second-order factor analysis on the eight sub-scales of which three performed both first-order and second-order factor analyses [25, 30, 31].

Table 1 shows how many papers satisfied the various methodological criteria.

A. Choice and justification of methods

1. Exploratory vs. confirmatory analysis

Perhaps the most striking result of this review is that in about half of the studies (15/28) the chosen method was inappropriate. In all these cases exploratory factor analysis was performed, whereas confirmatory analysis would have been more appropriate to answer the research question at issue. Both exploratory and confirmatory analyses were performed in only one study [33].

2. Exploratory analysis

Twenty-three studies performed exploratory analysis. Twenty studies used orthogonal rotation (15 used Varimax), for which no justification of the method was considered to be necessary. Three studies [25, 33, 34] used oblique rotation, one gave no justification for this choice [34]. The interpretation of the factor solution and its consequences was properly justified in 12/20 studies. In three studies [33, 35, 36], factor analysis was used to determine country-specific factor loadings to be

Table 2. General characteristics of the studies

Reference	SF-36 version	Population	Country	Number of patients included in FA	Aim of FA	FA on items or subscales
McHorney and Ware [7]	US english	Adults, visiting clinician	USA	3445	Confirmation of previously found factor structure	Subscales
Perneger et al. [43]	French	Young healthy adults (18–44 yr)	Switzerland	1007	Psychometric testing in new (healthy) population and language	Subscales
Wolinsky [32]	US english	Disadvantaged, older adults (> 50 yr) in clinic	USA	1051	Psychometric testing in new patient population	Items
Dexter et al. [25]	US english	Disadvantaged, older adults (> 50 yr) in clinic	USA	1053	Psychometric testing in new patient population	Items + Subscales
Essink-Bot et al. [47]	Dutch	Migraine patients (1); Matched control group (2)	Netherlands	496 (1), 515 (2)	Comparing the psychometric properties of four generic questionnaires	Subscales
Jenkinson et al. [37]	UK english	General population (18–64 yr)	UK	8204	Confirmation of previously found factor structure	Subscales
Lewin – Epstein et al. [29]	Hebrew	Jewish Israelies (45–75 yr)	Israel	2030	Psychometric testing in new language	Items
Mishra and Schofield [35]	Aus english	Representative women	Australia	41,546	Deriving Australian normative weights for calculating summary scores	Subscales
Reed [30]	US english	Adult health care workers	USA	394 + 461 (cross-validation)	Confirmation of previously found factor structure in new population	Items + Subscales
Stadnyk et al. [46]	UK english (acute version)	Frail elderly patients	Canada	146	Psychometric testing in new (elderly) population	Subscales
Sanson-Fisher and Perkins [44]	Aus english	General population	Australia	855	Psychometric testing in new language	Subscales
Apolone and Mosconi [39]	Italian	General population	Italy	2031	Psychometric testing in new language	Subscales
Fukuhara et al. [41]	Japanese	Industrial workers and their family	Japan	588	Psychometric testing in new language	Subscales
Fukuhara et al. [40]	Japanese	General population	Japan	3395	Psychometric testing in new language	Subscales
Ware et al. [9]	Danish, French, German, Italian, Dutch, Norwegian, Swedish, Spanish, UK english, US english	General population	Denmark, France, Germany, Italy, Netherlands, Norway, Sweden, Spain, UK, USA	4084, 3656, 2914, 1483, 1771, 2323, 8930, 9151, 2056, 2474	Equivalence testing of ten language versions	Subscales

Table 2. Continued

Reference	SF-36 version	Population	Country	Number of patients included in FA	Aim of FA	FA on items or subscales
Ware et al. [36]	Danish, French, German, Italian, Dutch, Norwegian, Swedish, Spanish, UK english, US english	General population	Denmark, France, Germany, Italy, Netherlands, Norway, Sweden, Spain, UK, USA	4084, 3656, 2914, 1483, 1771, 2323, 8930, 9151, 2056, 2474	Comparing country-specific versus standard scoring algorithms for calculating summary scores	Subscales
Keller et al. [28]	Danish, French, German, Italian, Dutch, Norwegian, Swedish, Spanish, UK english, US english	General population	Denmark, France, Germany, Italy, Netherlands, Norway, Sweden, Spain, UK, USA	4084, 3656, 2914, 1483, 1771, 2323, 8930, 9151, 2056, 2474	Confirmation of factor structure from one country in nine other countries	Subscales
Jenkinson et al. [38]	UK english (version II)	General population (18–64 yr)	UK	8889	Psychometric testing of new version (II)	Subscales
Chern et al. [49]	US english	Working population	USA	4255	Testing stability of factor structure over time	Subscales
Failde and Ramos [27]	Spanish	Patients with suspected ischemic cardiopathy (IC)	Spain	185	Psychometric testing in new patient population	Items
Scott and Haslett 1999 [26]	Australia/New Zealand adaptation	General population	New Zealand	7862	Psychometric testing in new general population	Subscales
Scott et al. [45]	?	General population of Maori (1), Pacific (2), and New Zealand Europeans (3)	New Zealand	1296(1), 168(2), 5467(3)	Psychometric testing in new (ethnic) populations	Subscales
Wilson et al. [33]	?	General population	Australia	18,492	Comparing scoring algorithms for calculating summary scores	Subscales
Goldbeck and Schmitz [48]	German	Patients with cystic fibrosis	Germany	70	Comparing the psychometric properties of three generic questionnaires	Subscales
Hardt et al. [34]	US english, UK english, German	patients with chronic fatigue syndrome	US, UK, Germany	740, 82, 65	Confirmation of previously found factor structure in new patient population	Subscales
Hobart et al. [42]	?	Adults with multiple sclerosis	UK	149(1), 288(2)	Psychometric testing in new patient population	Subscales
Thumboo et al. [31]	UK english (1), Chinese (Hong Kong) (2)	General population (21–65 yr)	Singapore	4122(1), 1381(2)	Psychometric testing in new (ethnic) populations	Items + Subscales
Thumboo et al. [50]	UK english and Chinese (Hong Kong) ^a	Bilingual volunteers	Singapore	157(2 FAs)	Equivalence testing of two language versions	Subscales

^a All subjects completed both versions.

used as weighing factors for the calculation of the physical and mental health subscale scores. Therefore, interpretation of the factor structure was considered to be irrelevant in these studies. Other goals of the factor analysis were confirmation of a previously found factor structure in five studies [7, 28, 30, 34, 37], psychometric testing of a new version [38] or in a new population or language in 15 studies [25–27, 29, 31, 32, 38–46], comparing the properties of different questionnaires [47, 48], and testing the stability of the factor structure over time [49]. Two studies [9, 50] performed equivalence testing of multiple language versions. Only three studies used a non-orthogonal factor structure [25, 33, 34], and two of those [25, 34] discussed the associations between factors.

3. *Confirmatory factor analysis*

Only six studies performed confirmatory factor analysis [28–30, 32, 33, 49]. They all adequately described the model to be confirmed and the strategy used to arrive at the best model. Five of these six studies interpreted the results correctly. However, only half of the studies (3/6) described the associations between the factors [28, 30, 49].

4. *Cross-validation*

Only four studies used a cross-validation design [28, 30, 41, 49], while in all 28 studies there were sufficient subjects to make cross-validation possible. None of these four studies randomly sub-divided their study population to perform cross-validation on the same data set. In three studies [28, 30, 41] the interpretation of the cross-validation results was convincing.

B. *Sample size and data quality*

1. *Sample size*

In most studies (25/28) the number of subjects was sufficient to perform factor analysis.

2. *Data quality: missing data*

Less than half of the studies (15/28) reported on missing data on the individual items. In one of these studies [49] the percentage of missing values raised concerns (25%) and no information was presented whether missings were missing at random. In eight studies [25, 29, 30, 33, 43, 46, 48, 50] missing data had been imputed by an acceptable method.

3. *Distributional properties*

The majority of studies reported the distribution of the items (21/28). Four studies reported distributional problems [26, 29, 30, 46]; one study handled these adequately [30].

C. *Full report of statistical entities [1]*

With respect to exploratory factor analysis, all studies stated whether they had performed principal component analysis or common factor analysis; criteria for retaining factors were presented in 16 of the 23 studies; eigenvalues and percentage of variance accounted for by the factors were presented in 14 of the 23 studies; only one study [26] failed to mention the rotation method. Three studies used a non-orthogonal method; two of these studies justified this method [25, 33] and also two studies [25, 34] presented the factor inter-correlations; almost all studies (20/23) presented the rotated factor loadings.

With respect to confirmatory factor analysis, all six studies mentioned the number of factors and the composition of the factors. Four studies stated whether they used orthogonal or correlated factors [28–30, 49] and three studies [28, 30, 32] described other model constraints. With respect to the fit, only one of the six studies [49] failed to mention the method of estimation. The overall fit was presented by all studies. Four studies also mentioned the relative fit [28, 30, 32, 49]; parsimony was discussed in only two studies [30, 49].

Five studies applied model modifications to improve the fit [28, 30, 32, 49]. Factor loadings were reported in four of these studies [28–30, 32, 49], and the communalities of the items or the squared correlations of observed variables with the factors in only one study [49]. Three studies reported correlations between the factors [28, 30, 49].

Discussion

As far as we know, this is the first critical assessment of factor analysis that has been performed in the field of health sciences. For this purpose, we developed a checklist, based on landmark papers on the use of factor analysis [1, 3, 4, 12, 14, 51]. Despite the pilot run to test the criteria list on five studies, during the assessment of the studies, more

adaptations to the rules had to be made before we were able to compile detailed instructions on how to decide on which issues were correctly reported and/or justified, and which methods were considered to be appropriate.

Floyd and Widaman [1] remarked that ‘the analytic technique used in a factor analysis appears to be the aspects of a factor analysis that are least understood by practicing scientists, and thus are often reported in insufficient detail in methods sections of journal articles’. Therefore their checklist focused on these technical aspects. In our opinion, the choices between the various possible methods of factor analyses are crucial, and therefore we added the section on choice and justification and the section on sample size and data quality. We felt the need to make a distinction between whether an issue was correctly reported and/or justified and whether an analysis was performed appropriately. Specific results needed to be reported. The reporting of methods is a prerequisite for assessing whether the right methods have been applied and to make replication of the study possible. In some cases we required a justification, especially when a number of arbitrary choices can be made, e.g. with regard to the number of factors to be rotated. If the authors described the methods they used, and justified their reasons for the choice of these methods, we were satisfied. In other situations we wanted to assess whether the method was appropriate, i.e. the method of factor analysis used. If the authors justified an exploratory analysis when a confirmatory method would have been appropriate for that specific research question, we assessed the adequateness of the method and not whether the authors did give a justification.

We acknowledge that the checklist we used included a number of criteria which required rather subjective judgement. Therefore, we describe the interpretation of the criteria in detail. Furthermore, in appendix A, we have presented the scores of all the papers that were included, to make these scores available and verifiable for readers.

With respect to the results, we were surprised by the fact that most authors performed exploratory factor analysis in situations in which a confirmatory factor analysis would have been more appropriate. We have been rather accommodating

in scoring this item. One can argue that in case of the SF-36 there is always a hypothesized factor structure. If the authors used an open formulation about exploring the factor structure in another situation we tolerated explorative factor analysis. The conclusion of an explorative factor analysis has far less power than confirmatory factor analysis to support a hypothesized factor structure [1]. This is nicely illustrated by the papers of Dexter et al. [25] and Wolinsky and Stump [32] who performed exploratory factor analysis and confirmatory factor analysis, respectively, on the same data-set. Dexter et al. [25] conclude that ‘In general, the principal component analysis confirmed the hypothesized underlying factor structure for each subscale and the overall SF-36’, while the factor loadings on the subscales General Health and Mental Health were, in our opinion, far from convincing. Wolinsky and Stump [32] conclude that ‘There is a problem, associated with the ‘getting sick’ and ‘getting worse’ items on the General Health subscale’ and ‘An alternative is to specify a nine-factor model that assigns ‘getting sick’ and ‘getting worse’ items to form their own construct, which is labeled ‘health optimism’’. They conclude that ‘A nine factor model including the separate construct of health optimism provides the best fit to the data’. It is recommendable to re-analyse studies by performing confirmatory analysis to determine whether the conclusions on the factor structure of the SF-36 would remain the same.

Other remarkable findings were that the interpretation of the final factor solution left much to be desired, and cross-validation was seldomly applied. With respect to the more technical aspects, the justification of which factors should be retained was often poor and the eigenvalues or percentages of variances accounted for by the factors was often not reported. Half of the studies in which confirmatory factor analysis was applied failed to report model constraints and factor loadings. Communalities of items were rarely mentioned.

As there are thousands of publications on factor analysis of instruments to measure health status, we had to restrict our selection in some way, either on the basis of specific journals, years of publication or specific instruments. This study focused on the SF-36. By choosing for the SF-36, we covered 13 different journals and the year of publication

varied from 1996 to 2002. The disadvantage of the choice for one instrument is that researchers may tend to imitate each other with regard to the methods used. We included five studies that performed factor analysis on the same data set [9, 28, 36] and [25, 32], respectively. As each research group can use a different strategy for their factor analyses (in both situations exploratory and confirmative factor analyses were used), we saw no reason to exclude these studies. We also included studies in which the same research groups analyzed different data sets: Jenkinson et al. [37, 38], Thumboo et al. [31, 50] and Ware and co-workers [7, 9, 28, 36, 39, 40]. In theory they may use different methods, depending on the specific purpose of the factor analysis. However, 9 of the 28 studies (32%) in this review were performed by members of the IQOLA group [5, 7, 9, 10, 28, 36, 39, 40, 43], who followed the methods as described and justified in the User's manual [8].

By choosing for the SF-36 as the subject of our methodological appraisal of factor analysis, we cannot ignore the validating work that has been done on the SF-36 and the methods that have been used by the developers. In the User's manual [8] they justify their choice for principal component analysis instead of common factor analysis and for orthogonal instead of oblique rotation. Advantages of principal component analysis over common factor analysis include (1) a simple additive model of factor content facilitating the interpretation of each scale; (2) summary measures that explain as much of the variance in the eight subscales as possible; (3) summary scales that are easy to estimate statistically; and (4) summary scores that are interpretable as physical and mental dimensions of health.

Advantages of orthogonal factor rotation over oblique factor rotation are that 'factor loadings', which are product moment correlations between scales and factors, can be squared and summed across factors to estimate the amount of variance in each scale accounted for by each factor, and the amount of variance in each scale is explained by all factors. As a result, factor content and implications for the interpretation of each scale are more straightforward.

We agree with the developers of the SF-36 and members of the IQOLA Project on these preferences for principal component analysis over com-

mon factor analysis and for orthogonal rotation instead of oblique rotation. However, we differ in opinion with regard to their choice for exploratory factor analysis instead of confirmatory factor analysis in some situations.

The choice of method of factor analysis is very much dependent on the purpose of the factor analysis. The aim of factor analysis may be (1) either pure data reduction; (2) assessment of the factor structure (dimensions) being measured by the questionnaire; (3) investigating whether the questionnaire shows the same dimensions across different groups (structural reliability); or (4) to determine country-specific factor loadings to be used as weighing factors for the calculation of the physical and mental health subscale scores.

For example, principal component analysis is adequate in the situation of development of an instrument (second aim), and also to determine factor loadings for the instrument to be used in another situation (fourth aim). However, in order to investigate whether the questionnaire shows the same (known) dimensions across different cultural populations or disease groups (third aim) confirmatory factor analysis is more appropriate than principal component analysis. Therefore we do not agree with the frequently cited sentence of Gandek and Ware [18] 'Principal component analysis, a type of factor analysis, was used in the IQOLA Project to estimate the congruence between hypothesized physical and mental health constructs and the SF-36 scales used to measure these constructs'. It is this situation which, in our opinion, requires confirmative factor analysis. Nowhere do they justify why they prefer to use exploratory factor analysis instead of confirmatory factor analysis for this purpose.

As many studies were done by members of the IQOLA project [10], the dependency of our results forms a limitation of our study. It has consequences for the generalisation of the results of our methodological assessment to other instruments: the type of shortcomings may be similar in other instruments, but the frequency of occurrence probably differs. Moreover, the generalisability of our findings probably varies per methodological criterion. Failure to use confirmatory factor analysis may be typical for the SF-36, because exploratory analysis is used by the IQOLA group. If we had extended our methodological assessment to all

health status questionnaires published in 'Quality of Life Research' in the past 5 years, we would have included more new measurement instruments, which required exploratory factor analysis because the factor structure is unknown yet.

The sub-optimal description of the method of factor analysis will probably also occur in the evaluation of other instruments.

This manuscript focussed on the quality of factor analysis performed on the SF-36. For that reason we ignored alternative methods that are available to examine the construct validity of health status questionnaires. Many validation studies on the SF-36 have been performed using other methods, such as the multitrait-multimethod approach [52, 53], and Rasch analysis [54]. Especially methods based on Item Response Theory, including estimation of differential item functioning, are powerful methods to check cross cultural equivalence of SF-36 translations [55] and the dimensionality of scales.

By identifying shortcomings in factor analysis and describing the specific points on which the studies fail to meet the criteria, we aim to improve the quality and reporting of future studies which perform factor analysis on the SF-36 or other health status questionnaires. To enable interested readers to make a full interpretation and replication of the results, the specific choice of analytic techniques requires justification, and the techniques used in any study should be described in sufficient detail. Moreover, as the interpretation of the results of factor analysis is rather subjective, it is important that a publication provides sufficient information to support the conclusions.

In analogy of the CONSORT statement [56] and the STARD statement [57, 58] to improve the quality of reporting of randomized clinical trials and diagnostic studies, respectively, journal editors, reviewers and authors could make use of our list of criteria to check on the completeness of reporting and the quality of performance of factor analysis, especially in studies in which factor analysis is the key element.

Conclusion

Ware [6] state that 'A major objective in constructing the SF-36 was achievement of high

psychometric standards. Guidelines for testing were derived from those recommended for use in validating psychological and educational measures by the American Psychological Association, the American Education Research Association and the National Council on Measurement in Education'. Despite the high standards used in the development of the SF-36, the quality of factor analysis in exploring or confirming the factor structure of the SF-36 leaves much to be desired. This critical assessment of methods of factor analysis applied in relation to the SF-36 showed that exploratory factor analysis is often used when confirmatory analysis would have been more appropriate and might have led to different conclusions. Furthermore, crucial information on the methods and the results of factor analysis is lacking in many publications. With this assessment we hope to contribute to the improvement of the application and reporting of factor analysis in health sciences research.

References

1. Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. *Psychol Assess* 1995; 7: 286-299.
2. Johnson DE. *Applied Multivariate Methods for Data Analysts*. 1998.
3. Streiner DL. Figuring out factors: The use and misuse of factor analysis. *Can J Psychiatry* 1994; 39: 135-140.
4. Gorsuch RL. *Factor Analysis*. Hillsdale NJ: Lawrence Erlbaum Associates, Inc, 1983.
5. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992; 30: 473-483.
6. Ware JE Jr, SF-36 health survey update. *Spine* 2000; 25: 3130-3139.
7. McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care* 1993; 31: 247-263.
8. Ware JE, Kosinski M, Keller SK. *Psychological and mental health summary scales: A user's manual*. 1994. Boston MA, The Health Institute (report).
9. Ware JE Jr, Kosinski M, Gandek B, et al. The factor structure of the SF-36 Health Survey in 10 countries: Results from the IQOLA Project. *International Quality of Life Assessment*. *J Clin Epidemiol* 1998; 51: 1159-1165.
10. Ware JE Jr, Gandek B. Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. *J Clin Epidemiol* 1998; 51: 903-912.

11. Ware J Jr, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. *Med Care* 1996; 34: 220–233.
12. Zwick WR, Velicer WF. A comparison of five rules for determining the number of components to retain. *Psychol Bull* 1986; 99: 432–442.
13. Allison DB, Gorman BS, Primavera LH. Some of the most common questions asked of statistical consultants: Our favorite responses and recommended readings. *Gene Soc Gen Psychol Monogr* 1993; 119: 153–185.
14. Bollen KA. *Structural Equations with Latent Variables*. New York: Wiley, 1989.
15. Pouwer F, Snoek FJ, Van der Ploeg HM, Ader HJ, Heine RJ. The well-being questionnaire: Evidence for a three-factor structure with 12 items (W-BQ12). *Psychol Med* 2000; 30: 455–462.
16. Kline P. *The Handbook of Psychological Testing*. London: Routledge, 1993.
17. Guadagnoli E, Velicer WF. Relation of sample size to the stability of component patterns. *Psychol Bull* 1988; 103: 265–275.
18. Gandek B, Ware JE Jr. Methods for validating and norming translations of health status questionnaires: The IQOLA Project approach. *International Quality of Life Assessment*. *J Clin Epidemiol* 1998; 51: 953–959.
19. Hays RD, Sherbourne CD, Mazel RM. The RAND 36-Item Health Survey 1.0. *Health Econ* 1993; 2: 217–227.
20. Kosinski M, Keller SD, Hatoum HT, Kong SX, Ware JE Jr. The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: Tests of data quality, scaling assumptions and score reliability. *Med Care* 1999; 37: MS10–MS22.
21. Reed PJ. Methods of evaluation in outcomes research. *Am J Manag Care* 1998; 4: 1616–1625.
22. Ware JE, Kosinski M. Interpreting SF-36 summary health measures: A response. *Qual Life Res* 2001; 10: 405–413.
23. Pfenning LE, Van der Ploeg HM, Cohen L, et al. A health-related quality of life questionnaire for multiple sclerosis patients. *Acta Neurol Scand* 1999; 100: 148–155.
24. Thumboo J, Feng PH, Soh CH, Boey ML, Thio S, Fong KY. Validation of a Chinese version of the Medical Outcomes Study Family and Marital Functioning Measures in patients with SLE. *Lupus* 2000; 9: 702–707.
25. Dexter PR, Stump TE, Tierney WM, Wolinsky FD. The psychometric properties of the SF-36 Health Survey among older adults in a clinical setting. *J Clin Geropsychol* 1996; 2: 225–237.
26. Scott KM, Haslett SJ. SF-36 health survey reliability, validity and norms for New Zealand. *Aust New Zeal J Public Health* 1999; 23: 401–406.
27. Failde I, Ramos I. Validity and reliability of the SF-36 Health Survey Questionnaire in patients with coronary artery disease. *J Clin Epidemiol* 2000; 53: 359–365.
28. Keller SD, Ware JE Jr, Bentler PM, et al. Use of structural equation modeling to test the construct validity of the SF-36 Health Survey in 10 countries: Results from the IQOLA Project. *International Quality of Life Assessment*. *J Clin Epidemiol* 1998; 51: 1179–1188.
29. Lewin-Epstein N, Sagiv-Schifter T, Shabtai EL, Shmueli A. Validation of the 36-item short-form Health Survey (Hebrew version) in the adult population of Israel. *Med Care* 1998; 36: 1361–1370.
30. Reed PJ. Medical outcomes study short form 36: Testing and cross-validating a second-order factorial structure for health system employees. *Health Serv Res* 1998; 33: 1361–1380.
31. Thumboo J, Fong KY, Machin D, et al. A community-based study of scaling assumptions and construct validity of the English (UK) and Chinese (HK) SF-36 in Singapore. *Qual Life Res* 2001; 10: 175–188.
32. Wolinsky FD, Stump TE. A measurement model of the Medical Outcomes Study 36-Item Short-Form Health Survey in a clinical sample of disadvantaged, older, black, and white men and women. *Med Care* 1996; 34: 537–548.
33. Wilson D, Parsons J, Tucker G. The SF-36 summary scales: Problems and solutions. *Soz Praventiv Med* 2000; 45: 239–246.
34. Hardt J, Buchwald D, Wilks D, Sharpe M, Nix WA, Egle UT. Health-related quality of life in patients with chronic fatigue syndrome: An international study. *J Psychosom Res* 2001; 51: 431–434.
35. Mishra G, Schofield MJ. Norms for the physical and mental health component summary scores of the SF-36 for young, middle-aged and older Australian women. *Qual Life Res* 1998; 7: 215–220.
36. Ware JE Jr, Gandek B, Kosinski M, et al. The equivalence of SF-36 summary health scores estimated using standard and country-specific algorithms in 10 countries: Results from the IQOLA Project. *International Quality of Life Assessment*. *J Clin Epidemiol* 1998; 51: 1167–1170.
37. Jenkinson C, Layte R. Development and testing of the UK SF-12 (short form health survey). *J Health Serv Res Policy* 1997; 2: 14–18.
38. Jenkinson C, Stewart-Brown S, Petersen S, Paice C. Assessment of the SF-36 version 2 in the United Kingdom. *J Epidemiol Commun Health* 1999; 53: 46–50.
39. Apolone G, Mosconi P. The Italian SF-36 Health Survey: Translation, validation and norming. *J Clin Epidemiol* 1998; 51: 1025–1036.
40. Fukuhara S, Ware JE Jr, Kosinski M, Wada S, Gandek B. Psychometric and clinical tests of validity of the Japanese SF-36 Health Survey. *J Clin Epidemiol* 1998b; 51: 1045–1053.
41. Fukuhara S, Bito S, Green J, Hsiao A, Kurokawa K. Translation, adaptation, and validation of the SF-36 Health Survey for use in Japan. *J Clin Epidemiol* 1998a; 51: 1037–1044.
42. Hobart J, Freeman J, Lamping D, Fitzpatrick R, Thompson A. The SF-36 in multiple sclerosis: Why basic assumptions must be tested. *J Neurol Neurosurg Psychiatry* 2001; 71: 363–370.
43. Perneger TV, Leplege A, Etter JF, Rougemont A. Validation of a French-language version of the MOS 36-Item Short Form Health Survey (SF-36) in young healthy adults. *J Clin Epidemiol* 1995; 48: 1051–1060.
44. Sanson-Fisher RW, Perkins JJ. Adaptation and validation of the SF-36 Health Survey for use in Australia. *J Clin Epidemiol* 1998; 51: 961–967.

45. Scott KM, Sarfati D, Tobias MI, Haslett SJ. A challenge to the cross-cultural validity of the SF-36 health survey: Factor structure in Maori, Pacific and New Zealand European ethnic groups. *Soc Sci Med* 2000; 51: 1655–1664.
46. Stadnyk K, Calder J, Rockwood K. Testing the measurement properties of the Short Form-36 Health Survey in a frail elderly population. *J Clin Epidemiol* 1998; 51: 827–835.
47. Essink-Bot ML, Krabbe PF, Bonsel GJ, Aaronson NK. An empirical comparison of four generic health status measures. The Nottingham Health Profile, the Medical Outcomes Study 36-item Short-Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument. *Med Care* 1997; 35: 522–537.
48. Goldbeck L, Schmitz TG. Comparison of three generic questionnaires measuring quality of life in adolescents and adults with cystic fibrosis: The 36-item short form health survey, the quality of life profile for chronic diseases, and the questions on life satisfaction. *Qual Life Res* 2001; 10: 23–36.
49. Chern JY, Wan TT, Pyles M. The stability of health status measurement (SF-36) in a working population. *J Outcome Meas* 2000; 4: 461–481.
50. Thumboo J, Fong KY, Chan SP, et al. The equivalence of English and Chinese SF-36 versions in bilingual Singapore Chinese. *Qual Life Res* 2002; 11: 495–503.
51. Gorsuch RL. Exploratory factor analysis: Its role in item analysis. *J Pers Assess* 1995; 68: 532–560.
52. Aaronson NK, Muller M, Cohen PD, et al. Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. *J Clin Epidemiol* 1998; 51: 1202.
53. Bjorner JB, Damsgaard MT, Watt T, Groenvold M. Tests of data quality, scaling assumptions, and reliability of the Danish SF-36. *J Clin Epidemiol* 1998; 51: 1001–1011.
54. Raczek AE, Ware JE, Bjorner JB, et al. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: Results from the IQOLA Project. International Quality of Life Assessment. *J Clin Epidemiol* 1998; 51: 1203–1214.
55. Bjorner JB, Kreiner S, Ware JE, Damsgaard MT, Bech P. Differential item functioning in the Danish translation of the SF-36. *J Clin Epidemiol* 1998; 51: 1189–1202.
56. Moher D, Schulz KF, Altman DG. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001; 357: 1191–1194.
57. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Ann Intern Med* 2003; 138: W1–12.
58. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med* 2003; 138: 40–44.

Address for correspondence: Henrica C. W. de Vet, Institute for Research in Extramural Medicine, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands
 Phone: +31-20-4448176; Fax: +31-20-4446775
 E-mail: HCW.deVet@vumc.nl

Appendix A. (Continued.)

	McH [7]	Perm [43]	Wol [32]	Dex [25]	Ess [47]	Jen97 [37]	Lew [29]	Mis [35]	Reed [30]	Stad [46]	San [44]	Apo [39]	Fuka [41]	Fukb [40]	Wa/ Ko [9]	Ga [36]	Kel [28]	Je99 [38]	Che [49]	Fai [27]	Sco [26]	Sco [45]	Wil [33]	Gold Har [48]	Hob [42]	Th01 [31]	Th02 [50]
C2.3			-			+		+	+								+		+				-				
C2.4			+			-		+	+								+		-				-				
C2.5			+			+		+	+								+		-				+				
C2.6			+			+		+	+								+		+				+				
C2.7			-			+		+	+								+		+				-				
C2.8								+	+								+		+								
C2.9			+			+		+	+								+		+								
C2.10			+			-		+	+								+		+				-				
C2.11			-			-		-	-								-		+				-				
C2.12			-			-		+	+								+		+				-				

+ - The description is both informative and methodologically sound.

- - The description is informative, but methodologically doubtful.

? - The description is too unclear or too incomplete to answer the question.

0 - No relevant information about this item is given in the paper.

blank - Not applicable.

¹ Smallest subsample was scored.² Largest subsample was scored.